DOCUMENT RESUME

ED 171 787 TM 009 258

AUTHOR Maxwell, Scott E.
TITLE Dependent Variable Reliability and Determination of Sample Size.
PUB DATE 11 Apr 79
NOTE 16p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, California, April 8-12, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Hypothesis Testing; *Mathematical Formulas; Predictor Variables; Reliability; *Research Design; *Sampling; Scores; Statistical Analysis
IDENTIFIERS *Sample Size; *Statistical Power; Test Length

ABSTRACT

Arguments have recently been put forth that standard textbook procedures for determining the sample size necessary to achieve a certain level of power in a completely randomized design are incorrect when the dependent variable is fallible because they ignore measurement error. In fact, however, there are several correct procedures, one of which is the standard textbook approach, because there are several ways of defining the magnitude of group differences. The standard formula is appropriate when group differences are defined relative to the within-group standard deviation of observed scores. Advantages and disadvantages of the various approaches are discussed, especially Subkoviak and Levin's alternative formula. Although all of the approaches discussed are mathematically correct, the traditional approach seems most meaningful; it is recommended that researchers continue to use this formula to determine statistical power and required sample size. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED171787

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Dependent Variable Reliability and Determination of Sample Size

Scott E. Maxwell

University of Houston

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Scott E. Maxwell

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Paper presented at 1979 AERA Annual Meeting, San Francisco,

April 11, 1979

Session 27.15

Please address all correspondence concerning this manuscript to:

Scott E. Maxwell
Department of Psychology
University of Houston
Houston, Texas 77004

TM009 258

Abstract

Arguments have recently been put forth that standard textbook procedures for determining the sample size necessary to achieve a certain level of power in a completely randomized design are incorrect when the dependent variable is fallible. In fact, however, there are several correct procedures, one of which is the standard textbook approach, because there are several ways of defining the magnitude of group differences. The standard formula is appropriate when group differences are defined relative to the within-group standard deviation of observed scores. Advantages and disadvantages of the various approaches are discussed.

One of the most frequently asked statistical questions in the social sciences is "How large must my sample be?" Implicit in this question is the recognition that the power of a statistical test must be considered, as well as the probability of a Type I error. The concept of power has received increasing attention in the social sciences in recent years (Brewer, 1972; Chase & Chase, 1976; Cohen, 1977; Schmidt, Hunter & Urry, 1976). Standard textbooks such as Glass and Stanley (1970), Hays (1973), Kirk (1968), and Winer (1971) contain extensive discussions of how to determine the appropriate sample size for a given experimental design.

In a series of articles, Levin and Subkoviak (1977, 1978) and Subkoviak and Levin (1977) have argued that the standard procedures advocated in textbooks are usually erroneous, even for a design as straightforward as a fixed effects completely randomized design. The basis for their argument is that earlier work (e.g., Cleary & Linn, 1969; Cleary, Linn, & Walster, 1970; Sutcliffe, 1958) showed that a decrease of error of measurement in the dependent variable is associated with increased statistical power, yet traditional formulas for determining sample size completely ignore error of measurement, despite the fact that almost all measures in the social sciences are fallible (i.e., they are not completely reliable). According to Subkoviak and Levin, traditional textbook formulas will underestimate the actual sample size necessary to obtain a certain level of power, with the degree of underestimation proportional to the reliability of the dependent measure. Hence, researchers following these formulas will take samples that are too small and will be too unlikely to reject the null hypothesis when a reasonable

alternative hypothesis is actually true.

How does Subkoviak and Levin's formula differ from textbook formulas? Traditional textbook formulas for determining power utilize an expression of the form

$$\phi = \sqrt{(n\psi^2) \; / \; (\nu+1) \; (\Sigma a_k^2)}, \qquad (1)$$

where ϕ is a noncentrality parameter, n is the sample size of each group, ν is numerator degrees of freedom (K-1 in the 1-way completely randomized design), and ψ represents a standardized linear combination of interest, i.e.,

$$\psi = (\Sigma a_k \mu_k) \; / \; \sigma \; . \qquad (2)$$

Actually, this notation corresponds to that used by Subkoviak and Levin, and differs somewhat from the notation used in most experimental design texts, where ϕ is defined as

$$\phi = \sqrt{n\Sigma(\mu_k - \mu)^2 \; / \; (\nu+1)\sigma^2} \qquad (3)$$

when equal sample sizes are assumed. Although formulas (1) and (3) appear to be different, it is possible to write formula (3) in terms of formulas (1) and (2) by letting

$$a_k = \mu_k - \mu \; . \qquad (4)$$

Algebraic manipulation shows that with this substitution into (1) and (2),

$$\psi^2 \; / \; \Sigma a_k^2 = \Sigma(\mu_k - \mu)^2 \; / \; \sigma^2 \; , \qquad (5)$$

so that formula (3) is actually a special case of formula (1). Indeed, this appears to have been Levin's (1975) rationale for the development of the more general formula for determining the power

of a particular comparison. In any event, Subkoviak and Levin's notation will be used here, with the understanding that the notation of formula (3) represents a special case.

Subkoviak and Levin maintain that for fallible measures it is necessary to revise formula (1) (and implicitly, formula (3)) to take into account the degree of unreliability of the dependent measure. They argue that the appropriate formula for the noncentrality parameter formula in this case is

$$\phi = \sqrt{(\rho n \psi^2) \; / \; (\nu+1) \; (\Sigma a_k^2)} \; , \qquad (6)$$

where ρ is the reliability of the dependent variable. The inclusion of ρ in their formula is the only apparent modification of the traditional formula. In fact, however, as Forsyth (1978) has pointed out, there is an additional difference that proves to be very important. The traditional definition of ψ (labeled ψ_X, for future reference) is

$$\psi_X = (\Sigma a_k \mu_k) \; / \; \sigma(X) \; , \qquad (7)$$

where $\sigma(X)$ is the within-group standard deviation of observed scores of the dependent measure. Subkoviak and Levin define ψ as

$$\psi_T = (\Sigma \; a_k \mu_k) \; / \; \sigma(T) \; , \qquad (8)$$

where σ (T) is the within-group standard deviation of true scores. Once this difference in defining ψ is taken into account, formula (6) is equivalent to formula (1). This can be seen by substituting

$$\rho = \sigma^2(T) \; / \; \sigma^2(X) \qquad (9)$$

for ρ and substituting formula (8) for ψ into equation (6), yielding equation (1) where ψ is defined as in formula (7). Thus, Subkoviak and Levin's formula is equivalent to the traditional formula; it is

not the case that one is correct and the other incorrect. The formulas differ only because of the different definitions of ψ. In addition, it is possible to derive other formulas that are equivalent to these two. For example, ϕ might be defined as

$$\phi = \sqrt{((1-\rho)n\psi^2) \; / \; (v+1) \; (\Sigma a_k^2)} \qquad (10)$$

defining ψ as

$$\psi_E = (\Sigma a_k \mu_k) \; / \; \sigma(E) \qquad (11)$$

where $\sigma(E)$ is the within-group standard deviation of error scores.

Which of these formulas is the appropriate one? According to Levin and Subkoviak, their formula (formula (6) here) incorporating the reliability of the dependent measure should be used, because formula (1) results in "underestimates of required sample sizes (or overestimates of available power)" (1977, p. 332). Forsyth, on the other hand, concluded that to use formula (6), "data analyses would have to be performed on true scores rather than observed scores" (1978, p. 380). Since analyses are actually carried out on observed scores, Forsyth concluded that formula (1) should be used. However, both Subkoviak and Levin and Forsyth are mistaken, because as previously shown, the formulas are all equivalent. Thus, it is not the case that one formula is correct and the others incorrect, as these authors have improperly concluded. If all of the parameters needed to compute ϕ from a formula are known, the formulas will produce equivalent results. The only difference between the formulas arises from the manner in which ψ is defined, i.e., how one wishes to describe the magnitude of the treatment effect -- in terms of the observed score standard deviation, or in terms of the true score standard deviation, or even in terms of the error score standard deviation.

Since the various formulas are all equivalent, it might seem that the choice of how to describe the magnitude of the treatment effect would not matter. However, this choice is in fact very important, both from a practical and a theoretical viewpoint, because if ψ_X equals some value and if formulas (1), (6), and (10) are to yield the same value of ϕ, ψ_T and ψ_E will not equal the value given by ψ_X, for a fallible dependent variable. For example, if $\rho = 0.64$, and if $\psi_X = 1$, then equivalent values for ψ_T and ψ_E are 1.25 and 1.67, respectively. Substitution of these values (as well as ρ or $1-\rho$ where needed) into formulas (1), (6), and (10), respectively shows that they all reduce to

$$\phi = \sqrt{n/(\nu+1)\ (\Sigma a_k^2)}. \qquad (12)$$

Thus, formulas (1), (6), and (10) are equivalent, but the attainment of numerically equivalent values of ϕ for the three formulas necessitates numerically different values of ψ, whenever $0<\rho<1$. The real issue, then, is not which formula is correct, but rather which definition of ψ is most appropriate, since stating that a meaningful alternative hypothesis for which power should be 0.80 is one corresponding to a ψ of 1 (for example) will in general yield different results for determining sample size, depending upon whether the ψ in question refers to ψ_X, ψ_T, or ψ_E. The reason that different results will be obtained is that a treatment effect of one standard deviation of observed scores is actually a larger effect than is an effect of one standard deviation of true scores, and hence requires fewer subjects for a fixed level of power. This relationship can be seen by inspecting formulas (1) and (6), and explains Subkoviak and Levin's conclusion that formula (1) underestimates the required number of subjects.

Which definition of ψ seems to be most generally appropriate? If a researcher wants to detect a treatment effect of "w standard deviation units" (where "w" is some constant) or of "w raw score units", should these units refer to observed score units, true score units, or error score units? Although any of the three can be used to yield correct results, it is difficult to imagine a situation where the use of error score units would be meaningful. Hence, in practice, the choice seems to be between the traditional approach versus Subkoviak and Levin's approach. Is it more meaningful to state magnitude of effect in terms of observed score units or in terms of true score units? Although the idea of using true score units sounds appealing, it would seem that the use of observed score units is actually more meaningful, for several reasons. First, use of Subkoviak and Levin's procedure requires the researcher to estimate the reliability of his/her measure. Although concern should be paid to reliability, accurate estimation may be difficult, and errors in estimation may produce misleading results. Second, and most important, it is possible that σ (T) might be near zero or even theoretically be zero yet group differences be large relative to $\sigma(X)$, because $\sigma(T)$ refers to within-group individual differences, and measures sensitive to group differences may be insensitive to individual differences (see Nicewander & Price, 1978). As an example, suppose that the dependent variable is a difference score whose components are highly correlated. In this situation it is well-known that the reliability of the dependent variable is quite low. In fact, as the correlation between the components increases, the reliability of the difference score approaches zero.

A researcher following formula (6), i.e., using Subkoviak and Levin's approach, will find that an enormous sample is required; indeed, as reliability approaches zero, the required sample size approaches infinity. However, Overall and Woodward (1977) have stated that the power to detect group differences is actually increased as reliability of the difference score decreases, contrary to the conclusion given by formula (6). The apparent contradiction arises because what may seem to be a reasonable magnitude of treatment effect in terms of true score units reflects a miniscule effect in terms of observed score units, since σ (T) is much smaller than σ (X). This relationship will occur anytime reliable individual differences within-groups are small; in this situation, Subkoviak and Levin's approach is unduly pessimistic. There seems to be no reason that the magnitude of between-group effects should be expressed in terms of reliable within-group individual differences variance, since the latter can conceivably be zero yet treatment effects exist. The problem seems to be that some researchers mistakenly believe that such a situation of no true variance within-groups necessarily implies that any between-group differences are due to error and hence are not real. However, this is incorrect, as the example of difference scores shows, so that it is sensible to express magnitude of treatment effect in terms of the observed within-groups standard deviation. Traditional textbook formulas are not only appropriate, but are also generally to be preferred to Subkoviak and Levin's formulation.

Since traditional textbook formulas are appropriate and do not take reliability of the dependent measure into account, it might seem that reliability has no effect whatsoever upon power. In particular,

it might seem that on the basis of the preceding discussion, lengthening a psychological test being used as the dependent variable in a study should have no effect on the power of the statistical test, although the reliability of the psychological test would be increased. Under certain reasonable assumptions, however, the power of the statistical test is increased by increasing test length. Consider a set of I parallel items with a common intercorrelation of ρ and a common within-group variance of σ_I^2. It will further be assumed that the magnitude of the treatment effect is a constant for each item, i.e., there is a constant C such that for every item

$$\sum_K a_K \mu_K = C. \tag{13}$$

Thus, for the test composed of I items,

$$\sum_i^I \sum_K a_K \mu_K = IC, \tag{14}$$

i.e., the magnitude of the effect for the total test score is I times C. The within-group variance of test scores is given by

$$I\,\sigma_I^2\,(1 + (I-1)\,\rho), \tag{15}$$

since test scores are simply sums of scores on individual items. Substituting these expressions into formula (1) for ϕ yields

$$\phi = \sqrt{nI^2C^2/I\sigma_I^2(1 + (I-1)\rho)\,(\nu+1)\,(\Sigma a_K^2)} \tag{16}$$

since for the test as a whole,

$$\psi = IC/\sqrt{I\sigma_I^2\,(1 + (I-1)\rho)} \tag{17}$$

Simplification of (16) leads to

$$\hat{\phi} = \sqrt{nIC^2/\sigma_I^2(1 + (I-1)\rho)\,(\nu+1)\,(\Sigma a_K^2)}. \tag{18}$$

This formula is interesting for several reasons. First, it provides yet another approach for determining power. Once again, the

distinction between this formula and the others arises because of the definition of ψ. With formula (18), ψ is expressed in terms of the magnitude of treatment effect for the observed score standard deviation of a single item. Second, unless $\rho=1$ (which corresponds to an infallible measure), ϕ increases with increasing I, so that lengthening the dependent measure results in increased statistical power. Third, the formula shows the relative effects on power of lengthening the test and increasing the sample size. Using this formula it is possible to estimate the gain in power that would result from lengthening the test by some number of items and/or increasing the sample size by some number of subjects. In practice, the formula should be viewed as providing an estimation, since the assumption of parallel items may be violated, and since the researcher's judgments of C, ρ, and σ_I^2 will probably be subject to error. Fourth, the role of ρ in affecting the extent to which lengthening the test increases power is made explicit. For example, if $\rho=1$, formula (18) simplifies to

$$\phi = \sqrt{nC^2/\sigma_I^2 \ (\nu+1) \ (\Sigma a_K^2)} \ , \qquad (19$$

so that increasing the length of the test has no effect on power. Such an occurrence is reasonable because ρ will equal 1 if and only if the dependent measure is infallible. Thus, for an infallible variable, length of test is irrelevant. At the other extreme, if $\rho=0$, the formula simplifies to

$$\phi = \sqrt{nIC^2/\sigma_I^2 \ (\nu+1) \ (\Sigma a_K^2)} \ , \qquad (20)$$

so that in terms of ϕ there is a direct trade-off between number of items and number of subjects. However, number of subjects is actually more influential even here, since the degrees of freedom for the

denominator depends upon sample size but not number of items, and larger denominator df is associated with higher power.

At this point, the fact that length of test influences power with a fallible dependent measure may seem to contradict the fact that traditional formulas that ignore reliability are appropriate for determining power. Once again the apparent paradox is a consequence of different definitions of ψ. If magnitude of effect is expressed in terms of the observed standard deviation of a single item, increasing the number of items results in increased power. However, if magnitude of effect is expressed in terms of observed standard deviation of the entire test, lengthening the test has no effect on power, because the standard deviation of observed scores increases as the test becomes longer. In this sense, a 1 standard deviation difference between groups on a 20-item test means that the groups are more different from one another than would a 1 standard deviation on a 40-item test. Specifically, the magnitude of the effect for an individual item would be larger for the shorter test than for the longer test. Hence, a researcher who uses a short test and specifies a ψ to be used in formula (1) is actually expressing a larger treatment effect per item than is another researcher who specifies the same value for ψ but uses a longer test. Although formula (1) would tell the two researchers that for a fixed n their power is the same, the researcher with the shorter test must actually have a larger treatment effect per item to achieve the same level of power as the researcher using the longer test. Again, formulas (1) and (18) are mathematically equivalent, and differ only in the way in which ψ is defined.

As has been demonstrated, there are numerous correct methods for determining power. The choice between these methods hinges upon how one wishes to express the magnitude of treatment effects. In other words, it is possible to state an alternative hypothesis against which power is to be judged in several different ways, and power varies with these different alternative hypotheses. A treatment effect of 1 standard deviation of observed scores may represent an effect of only 0.1 standard deviation of true scores, so that whether such an effect is considered to be relatively large or quite small depends upon the frame of reference. Although all of the approaches are mathematically correct, the traditional approach seems most meaningful, and it is recommended that researchers continue to use this formula to determine power and required sample size.

References

Brewer, J. K. On the power of statistical tests in the American Educational Research Journal. American Educational Research Journal, 1972, 9, 391-401.

Chase, L. J., & Chase, R. B. A statistical power analysis of applied psychological research. Journal of Applied Psychology, 1976, 61, 234-237.

Cleary, T. A., & Linn, R. L. Error of measurement and the power of a statistical test. British Journal of Mathematical and Statistical Psychology, 1969, 22, 49-55.

Cleary, T. A., Linn, R. L., & Walster, G. W. Effect of reliability and validity on power of statistical tests. In E. F. Borgatta and G. W. Bohrnstedt (Eds.), Sociological methodology. San Francisco: Jossey-Bass, 1970.

Cohen, J. Statistical power analysis for the behavioral sciences. New York: Academic Press, 1977.

Forsyth, R. A. A note on "Planning an experiment in the company of measurement error" by Levin and Subkoviak. Applied Psychological Measurement, 1978, 2, 379-383.

Glass, G. V., & Stanley, J. C. Statistical methods in education and psychology. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Hays, W. L. Statistics for the social sciences. New York: Holt, Rinehart, and Winston, 1973.

Kirk, R. E. Experimental design: Procedures for the behavioral sciences. Belmont, CA.: Brooks/Cole, 1968.

Levin, J. R. Determining sample size for planned and post hoc analysis of variance comparisons. Journal of Educational Measurement, 1975, 12, 99-108.

Levin, J. R., & Subkoviak, M. J. Planning an experiment in the company of measurement error. Applied Psychological Measurement, 1977, 1, 331-338.

Levin, J. R., & Subkoviak, M. J. Correcting "Planning an experiment in the company of measurement error." Applied Psychological Measurement, 1978, 2, 384-387.

Nicewander, W. A., & Price, J. M. Dependent variable reliability and the power of significance tests. Psychological Bulletin, 1978, 85, 405-409.

Overall, J. E., & Woodward, J. A. Unreliability of difference scores: A paradox for measurement of change. Psychological Bulletin, 1975, 82, 85-86.

Schmidt, F. L., Hunter, J. E., & Urry, V. W. Statistical power in criterion-related validation studies. Journal of Applied Psychology, 1976, 61, 473-485.

Subkoviak, M. J., & Levin, J. R. Fallibility of measurement and the power of a statistical test. Journal of Educational Measurement, 1977, 14, 47-52.

Sutcliffe, J. P. Error of measurement and the sensitivity of a test of significance. Psychometrika, 1958, 23, 9-17.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1971.